Case Study: Centre for e-Research (CeRch), King's College London

Primary contact for case study: Gareth Knight, Digital Curation Specialist

Additional input from: Richard Palmer, System Administrator

1. Overview

The general aim of the case study is to explore the real-world potential for implementing identified Greening Information Management (GIM) methods. Each case study undertaken will determine current information management practice across a specific information service/collection within Higher Education Institutions (HEIs). It will then assess the feasibility of implementing GIM methods within that environment and consider the costs and benefits to the organisation as a result of such implementation(s).

2. Introduction

The Centre for e-Research (CeRch) is located in Information Services and Systems at King's College London with a broad remit to work across disciplinary areas. CeRch works at the intersection between cross-disciplinary research and practice in information creation and management, knowledge production and ICT methods, tools and technologies. The Centre incorporates the former AHDS Executive and the Methods Network.

3. Phase 1: Examining the current IE

CeRch currently holds the following types of information and resources:

1. Legacy AHDS collections
    a. Performing Arts
    b. Literature, Languages and Linguistics
    c. History
    d. Visual Arts
2. JISC-funded projects
    a. East London Theatre Archive (ELTA)
    b. Stormont Papers
    c. Historical Hansard (forthcoming)
3. JISC collections
    a. Proquest
    b. Brill Journal Archive
    c. Early English Books Online
    d. Others

3.1 Stewardship requirements

Different stewardship requirements apply to each of the types of information or resources held, as follows:

1. Legacy AHDS collections
   CeRch does not have a formal commitment to curate and preserve digital collections since April 2009, since the closure of the AHDS. However, it continues to maintain the web site and provide limited curation for collections. Legacy AHDS collections are stored as a demonstration of good will to the arts and humanities research community. In many cases (e.g. the performing arts collections) the digital collections are not made available elsewhere. The continued provision of service for 10+ years demonstrates our expertise in providing long-term digital curation.

2. JISC-funded projects
   JISC imposes a commitment to maintain collections as a requirement of funding. For most collections, JISC specifies a standard lifetime of project + 3 years clause. For the ELTA collection, CeRch has committed to making it available for 10 years. However, the aim is to provide access to these collections in the longer-term.

3. JISC collections
   CeRch has a contractual obligation to provide bitstream preservation and content preservation within a "dark archive", i.e. no public access provision.

Formal assessment of CeRch's information environment has been undertaken using the DRAMBORA toolkit to assess the broad risks associated with data storage and management across all digital collections and projects. This has informed the development of management practices and procedures. DAF is deemed less applicable since a good understanding of the data held, where it is stored and how it should be managed already exists.

The Preservation Exemplars at Kings (PEKin) project at KCL is currently using a combination of DAF and DRAMBORA to assess digital assets within the college.

3.2 Management of resources

CeRch employs a small team (Digital Curation Manager, Digital Curation Specialist and an Information & Management Specialist) who is responsible for curation and preservation related issues within the department. Digital resources created or captured by the centre are managed by members of the project team and a set of procedures are followed to ingest data collections, store them on a preservation system and, if possible, make it available to a designated community. Data is stored on two Dell RAID arrays (RAID 5) located at two distinct locations. The on-site system provides 15 terabytes of storage, while the off-site system currently has a

capacity of 9 terabytes. Both systems are managed through a Debian Linux installation running in a VMWare virtual machine.

At the time of writing, digital resources are stored and managed using a file system directory hierarchy. The set of procedures to be followed for managing digital collections, including the creation of an OAIS Archival Information Package (AIP) and Dissemination Information Package (DIP) are established in an Ingest Manual and a set of preservation policy documents. In the near future, it is planned that the management process will become largely automated, through the implementation of a Fedora-based repository and the adoption of automated workflow tools.

The management system operated by the Centre maps on to the Open Archive Information System (OAIS) Reference Model and conforms to most of the criteria specified by TRAC (data management procedures are being reworked to address some of the missing elements). KCL uses various technical standards as a subset of their management operation.

3.3 Rationalisation of resources

It is currently uncertain whether resources/files might be rationalised in some way. Although there is no legal requirement to maintain the AHDS data, there may be implications for the organisational reputation if it were rationalised. It is possible that file content might potentially be rationalised through means including intra-file de-duplication, although this is not certain.

4. Phase 2: Evaluating techniques to green IM

Three techniques of a list of seven presented were deemed relevant to CeRch. These are tiered storage, the use of a storage repository and de-duplication.

4.1 Tiered storage

CeRch already implement a tiered storage model to a degree. For example, one collection of resources is held on a dark archive and material is drawn from that upon request. Other collections are made available continuously without restriction and are therefore held on high performance servers. The Centre plans to make greater use of this type of tiered approach to collection management. It is probable that the full range of information held by the Centre could be handled in this way.

4.1.1 Local benefits

Savings can be made by varying the specification of hardware required to host different aspects of the overall collection. This model would enable an organisation to prioritise collections for access and lower grade hardware, and less power, would therefore be needed for less frequently used collections. For example, the majority of AHDS collections are ideally made available 24/7 on a fully open access basis, whereas less actively used collections such as JISC collections may be held on less

readily accessible platforms such as lower specification (and ideally lower power consumption) servers or even on offline on tape storage facilities.

A tiered storage setup could facilitate the potential to audit collections being held. For example, when resource availability periods specified by funders are exceeded, collections being held could be evaluated strategically, to decide what level of priority should be given to their storage and access. This process would improve general information retention and disposal within the Centre.

4.1.2 Local drawbacks

There is a policy in place to store files uncompressed and this means that storage within lower-priority hardware cannot be optimised locally. Although not a drawback as a result of the adoption of the tiered storage approach, this policy could potentially limit the extent of the benefits achieved using the approach.

4.1.3 Institutional benefits

The increased ability to prioritise the use of high-performance hardware is likely to result in overall energy savings at an institutional level. Institutional stewardship is also likely to benefit from this approach.

4.2 De-duplication

Intra-file de-duplication (e.g. where a common corporate image is used within a large number of files, this image can be stored once only with links automatically inserted to point to the image from individual files) is a potential means of reducing the disk storage used by CeRch. Within the AHDS collections, a small amount of duplication occurs, primarily in web site content. XCDL and XCEL have already been considered as means of identifying differences and similarities between files held by the Centre.

The Centre for e-Research, in its role as the AHDS Executive adopted a data management strategy that complied with the OAIS Reference Model, storing multiple manifestations of the same object on disk. Each collection was separated into 2-3 directories that contained an Submission Information Package (SIP) as provided by a depositor, an Archival Information Package (AIP) that represented a preservation master and, in most circumstances, an Dissemination Information Package (DIP) for distribution to a user. A technique that could limit this type of duplication (e.g. a 'migration on demand' service, as proposed in the CEDARS project) would likely reduce storage requirements, at the expense of increasing processing requirements to produce derivatives on-the-fly.

The applicability of de-duplication techniques to all types of information held by CeRch is dependent on the policies imposed by, and the contracts held with, information providers and funders. For example, JISC Collections and JISC projects could not be subjected to any de-duplication techniques introduced.

4.2.1 Local benefits

The introduction of de-duplication techniques would assist in the identification of effective long-term storage requirements, which would further the business of the centre whose primary activity is the preservation of information.

The technique would ultimately result in storage savings, leading to freed up hardware, leading to financial savings.

4.2.3 Institutional benefits

Financial and greening costs are likely to be achieved owing to a reduction in the amount of hardware required and energy consumed.

4.3 Storage repository

CeRch has plans to implement Fedora (http://www.fedora-commons.org/) to house a number of its collections. Fedora will be used to manage all the different manifestations of an object. Various automated systems will be introduced to handle metadata creation, format conversion, and so on.

Fedora has been used as the back-end for projects like the East London Theatre Archive (ELTA). Files are stored within Fedora, with a second system being used to serve them up. Another service, SOAPI (Service-Orientated Architecture for Preservation and Ingest)[1], has developed processes to ingest data to Fedora. Each project will have a distinct output, which can potentially be ingested to a single instantiation of Fedora. Content models can be developed within Fedora to store collections as required.

Gareth Knight is involved in another project – the aforementioned PEKin (Preservation Exemplars at King's)[2] – which is developing a repository for research data and administrative records. It investigates the information lifecycle of research and administrative data. Although these data types differ in terms of the length of time for which they must be stored and the activities that they must support, management of the data lifecycle requires the application of similar processes. A common infrastructure to manage the lifecycle of such resources is being developed. DAF has been used to inform a new data assessment methodology to help ascertain

[1] http://www.kcl.ac.uk/iss/cerch/projects/completed/soapi.html
[2] http://www.kcl.ac.uk/iss/cerch/projects/portfolio/pekin.html

(e.g. where data is stored) who is responsible for it. This has been combined with the risk assessment element of DRAMBORA, to provide a robust assessment framework. Based on requirements and risks, decisions can be made on when resources should be moved to an alternative storage platform, and what nature this platform should take. PEKin is due to complete in October 2010.

4.3.1 Local benefits

Tasks including metadata creation and format conversion have previously been handled manually i.e. for the AHDS collections. Fedora enables these types of processes to be automated, saving staff time and increasing consistency.

It is desirable to store resources in different locations. Fedora enables the identification of an individual data stream, or record; the submission version of a record can be stored on a slower storage area; the archival and dissemination versions can be stored on a higher-performance server. Fedora then facilitates the establishment of relationships across these various manifestations.

The use of Fedora (or indeed other repository software) has the potential to reduce the storage requirements through the ability to create relationships between related digital objects. The development of Fedora disseminators would enable derivatives to be generated on-the-fly (e.g. the creation of a JPEG derivative from a TIFF image for dissemination)[3]. The use of Fedora is likely to increase effective stewardship due to the introduction of automated processes, previously undertaken on a manual basis. An example would be the increased efficiency in the consistent creation of preservation metadata, if undertaken on an automated basis.

4.3.3 Institutional benefits

The effective demonstration of long-term data management engenders trust in the institution from external parties. Processes embedded within the repository setup can potentially further this.

The use of Fedora is likely to enable more efficient compliance with legislation. Increased automation makes it easier to demonstrate that a set of procedures have been followed. PLANETS has looked at experimental workflows as they are linked to repositories, which may provide a better insight to this.

4.3.4 Institutional drawbacks

The development of content models and disseminators tailored to the requirements of arts and humanities data requires development in the early stages of repository introduction, which may incur up-front costs. The ingest of complex digital

[3] Further details of Fedora Disseminators are available at http://www.fedora-commons.org/download/2.2/userdocs/tutorials/tutorial2.pdf

resources may require manual ingest to model the inter-relationship between data objects.

5. Phase 3: Assessing costs and benefits

The use of a data repository, more specifically Fedora, was selected as the choice of GIM technique considered most appropriate for CeRch.

5.1 Working practices

Various changes to working practices are likely to be required, depending on how Fedora is implemented. For example, a deposit interface can be created so that third parties can contribute data, which would require to be overseen. The approach makes for rapid identification of common anomalies, for example, an inappropriate image format being uploaded. An individual would be required to be in place to resolve unexpected issues and to exert quality control and consistency across deposits.

Further, depending on the extent of collections being uploaded, scalability will become an issue, which will inform whether specific working practices should be manual or automated.

5.2 Business implications

A common infrastructure could potentially be applied to many different projects, thereby optimising funding and streamlining working practices and processes. The increased ability to automate procedures as a result would have a direct impact on staff time.

In turn, training requirements would be reduced, overheads are likely to be reduced, and users of the service are likely to develop a perspective of trust. Guaranteed compliance with a certain set of criteria will enhance the organisation's reputation.

5.3 Digital footprint

A centralised management system would facilitate the distribution of storage systems (and hardware), thought likely to result in a reduction in the amount of energy consumed. The use of Grid Bricks[4][5] is being considered to distribute certain aspects of storage and processing between multiple low powered machines. Each 'brick' is represented by a Mini-ITX system containing minimal hardware. The

[4] For a paper on Grid Bricks: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1194830

[5] Further detail is available at http://portal.acm.org/citation.cfm?id=825010.

iRODS (Integrated Rule-Oriented Data System) middleware will be used to provide data management across heterogeneous devices.

The Grid Bricks approach is likely to have an impact on the lifecycle of hardware and its management. It will be cheaper to replace one low-powered machine as it fails or becomes obsolete, in comparison to the cost of replacing a larger, more costly server. Hardware lifecycle could therefore become shorter. For example, five low powered machines could be replaced reasonably regularly (e.g. every 3-4 years). In contrast, a single, expensive server would need to demonstrate value over a longer timescale (e.g. every 7 years). This would also reduce reliance upon a single supplier or technology, which has the potential for an institution or department to choose alternative hardware that is more competitively priced.

There will also be an impact on the lifecycle of hardware and its management. It will be cheaper to replace one low-powered machine as it fails or becomes obsolete, in comparison to the cost of replacing a larger, more costly server. Hardware lifecycle could therefore become more cyclical. For example, five low powered machines could be replaced reasonably regularly (e.g. every 3-4 years). In contrast, a single, expensive server would need to demonstrate value over a longer timescale (e.g. every 7 years). This would also reduce reliance upon a single supplier or technology, which has the potential for an institution or department to choose alternative hardware that is more competitively priced.

5.4 Change management

The process of introducing Fedora would bring implications for managing associated processes. Considerable changes would be required in terms of staff training, infrastructures operated and so on. CeRch are introducing it on a gradual basis, to try to identify all the areas of change that are likely to require management.

Different types of staff would need to be employed. More development staff would be required in the short term; this may then change once the system was well-established within the institution. Allocation of staff time and resources, and therefore training, would also be subject to change.

5.5 Evaluating costs/benefits

Benefits are likely to be found in relation to staff time, although the level of benefit will vary depending on the degree of implementation time required, and the level of customisation desired. Externally generated benefits to the organisation would include positive client perception.

There would be additional costs in relation to training users in the use of the new system.

Systems support requirements will change, although it is unclear whether this will constitute an overall cost or benefit. For example, administrative staff may be able to manage the system the majority of the time, with a more costly system developer contributing to the management perhaps one day per week, as opposed to full time.

Where a benefit to one stakeholder constitutes a cost to another, prioritisation would vary from case to case. CeRch has a specific task allocation, determined by users. User needs, funding requirements and so on require balancing against costs and benefits, within a certain set level of resourcing.

The introduction of a new system (e.g. like PEKin currently being introduced within the College archive section) could bring benefits to the institution as a whole. For example, should a committee or particular user group fail to submit papers, this can be easily identified through a process of regular review.

Potential barriers to the introduction of Fedora include lack of interoperability with other existing and established services; lack of general acceptance of open source products; uncertainty of maintenance costs and requirements of in-house systems when compared to commercial products; lack of a business case for green computing techniques e.g. moving away from a monolithic server to many different greener systems, potentially capable of combining their processing cycles; lack of business justification.

The success of Fedora's implementation could be measured or evaluated based upon the functionality offered for services via the new platform. Cost implications compared with an existing system, or outsourced system, could be quantified. It is deemed difficult to identify such measures e.g. level of user satisfaction. It is straightforward to identify what the requirements are, and that they have been met, but how can this be expressed in quantitative terms? The strategic development of the technical infrastructure would be within the remit of the Deputy Director/Technical Manager of CeRch. Evaluation would be performed by the System Manager and digital curation team.